

A Data Science Approach of

Similarities and Differences Between Chicago Communities and Their Relationship with Community Average Income

Peihong Man



Part 1: Introduction/Business Problem

First let's consider this question. Imagine in some city, if someone is looking to open a restaurant, where would you recommend that they open it? Similarly, if a contractor is trying to start their own business, where would you recommend that they setup their office?

By first looking at it, this could be a very complicated problem. One needs to do a lot of research, get familiar with the city, find relevant information's online, drive around ask local people, etc. But what if you want an easier method to get it done, which can make sure you don't miss a spot? This is where data science kicks in.

In this project, we will take the city of Chicago as an example. We will first get all the community areas and neighborhoods in the city of Chicago, then we will get the coordinates of them. After that, we will use the API from Foursquare to extract the most visited venues in each community area. Then we will use unsupervised machine learning method - k-means to divide the communities into several clusters. You will see the similarities and differences between them. Last but not least, we will do some analysis on the data, and use the income data from Chicago Data Portal to help us figure out the deep relationship between popular venues and communities' characters.

Part 2: Get Our Dataset No. 1

In this part, I will use the Foursquare location data to solve the problem and execute my idea. I will also use other datasets in combination with the Foursquare location data, which will be explained in the next part.

First we need to get all neighborhoods in Chicago. I used BeautifulSoup to scrape data from wiki page. https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago. The page looks like this:

List of neighborhoods in Chicago

From Wikipedia, the free encyclopedia

There are sometimes said to be more than 200 **neighborhoods in Chicago**,^[1] though few residents would agree on their names and boundaries. A city ordinance prescribing and mapping 178 neighborhoods^[2] is almost unknown and ignored even by municipal departments. **Neighborhood** names and identities have evolved over time due to **real estate development** and changing **demographics**.^[3] The City of **Chicago** is also divided into 77 **community areas** which were drawn by **University of Chicago** researchers in the late 1920s.^[4] Chicago's community areas are well-defined, generally contain multiple neighborhoods, and are less commonly used by city residents.^{[3][5]} More historical images of Chicago neighborhoods can be found in **Explore Chicago Collections**, a digital repository made available by **Chicago Collections** archives, libraries and other cultural institutions in the city.^[6]

Contents [hide]

- 1 List of neighborhoods by community area
- 2 See also
- 3 References
- 4 External links

List of neighborhoods by community area [edit]

Neighborhood	Community area
Albany Park	Albany Park
Altgeld Gardens	Riverdale
Andersonville	Edgewater
Archer Heights	Archer Heights

After we scraped the data using BeautifulSoup, we process the raw data we have.

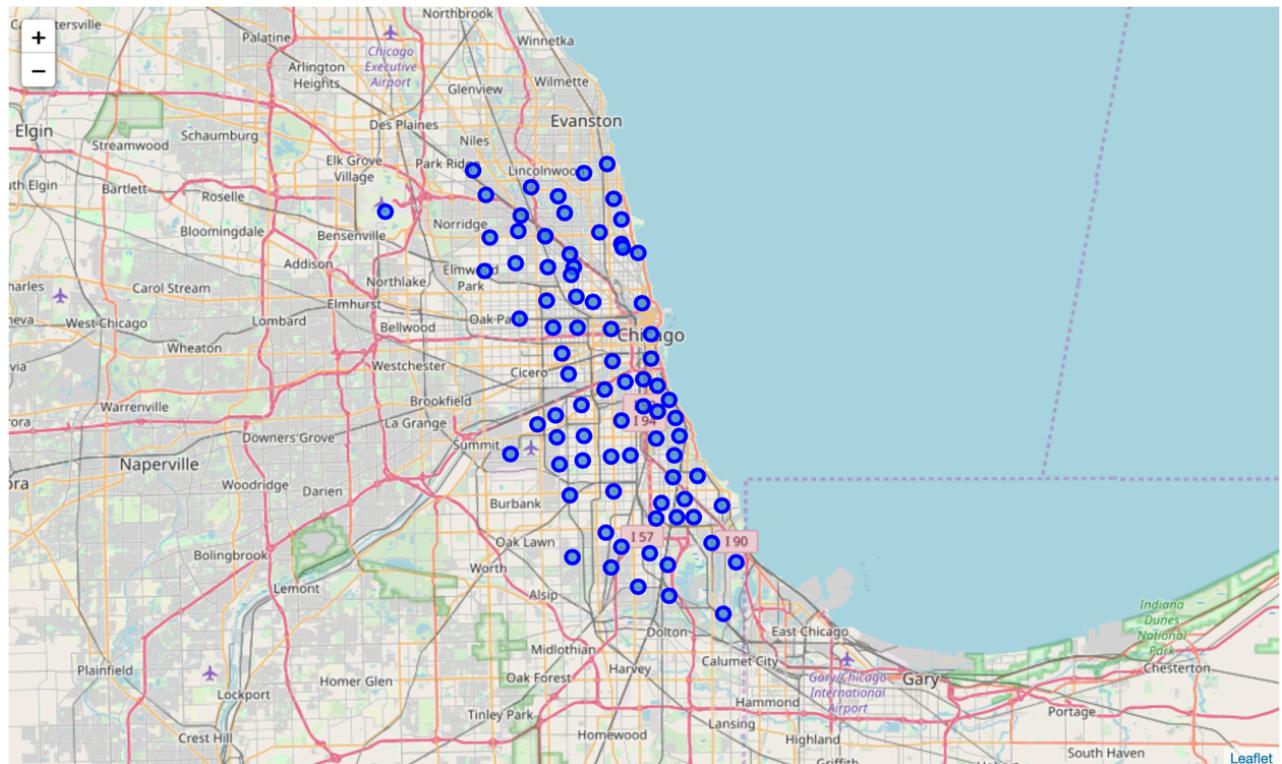
Here the processing steps:

- 1) Group the neighborhoods based on their community area.
- 2) Get the coordinates of all these community areas.
- 3) Manually add missing coordinates by simply searching them in google.
- 4) Corrected one error coordinate.

After the processing, here is what the dataframe looks like:

	CommunityArea	Neighborhood	Latitude	Longitude
0	Albany Park	Albany Park, Mayfair, North Mayfair, Ravenswo...	41.9719	-87.7162
1	Archer Heights	Archer Heights	41.8114	-87.7262
2	Armour Square	Armour Square, Chinatown, Wentworth Gardens	41.84	-87.6331
3	Ashburn	Ashburn, Ashburn Estates, Beverly View, Crest...	41.7475	-87.7112
4	Auburn Gresham	Auburn Gresham, Gresham	41.7505	-87.6643
5	Austin	Galewood, The Island, North Austin, South Austin	41.8879	-87.7649
6	Austin, Humboldt Park	West Humboldt Park	41.9025	-87.7361
7	Avalon Park	Avalon Park, Marynook, Stony Island Park	41.745	-87.5887
8	Avondale	Avondale, Jackowo, Wacławowo	41.9389	-87.7112
9	Avondale, Irving Park	Polish Village	41.9534	-87.7364

Now we have our data, let's generate a map of Chicago with all communities using folium.



We found totally 89 communities near Chicago area. They are evenly distributed.

Now we are ready to get all venues near each community. Here we use Foursquare database.

Below is the table listing 10 most common venues in each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Albany Park, Mayfair, North Mayfair, Ravenswooo...	Sandwich Place	Pizza Place	Donut Shop	Fried Chicken Joint	Bakery	Mobile Phone Shop	Gas Station	Diner	Chinese Restaurant	Bus Station
1	Andersonville, Edgewater, Edgewater Beach, Edg...	Bank	Asian Restaurant	Sandwich Place	Sushi Restaurant	Mexican Restaurant	Coffee Shop	Pharmacy	Antique Shop	Video Store	Mobile Phone Shop
2	Archer Heights	Mexican Restaurant	Mobile Phone Shop	Grocery Store	Gas Station	Optical Shop	Candy Store	Big Box Store	Bar	Bank	Bakery
3	Armour Square, Chinatown, Wentworth Gardens	Chinese Restaurant	Sports Bar	Hot Dog Joint	Breakfast Spot	Gas Station	Italian Restaurant	Asian Restaurant	Indian Restaurant	Sandwich Place	Grocery Store
4	Ashburn, Ashburn Estates, Beverly View, Crestl...	Construction & Landscaping	Cosmetics Shop	Light Rail Station	Italian Restaurant	Automotive Shop	Cuban Restaurant	Currency Exchange	Fish Market	Fish & Chips Shop	Filipino Restaurant

One can see that different neighborhoods have totally different common venues. One thing obvious is Chinatown does have a lot of Chinese restaurants.

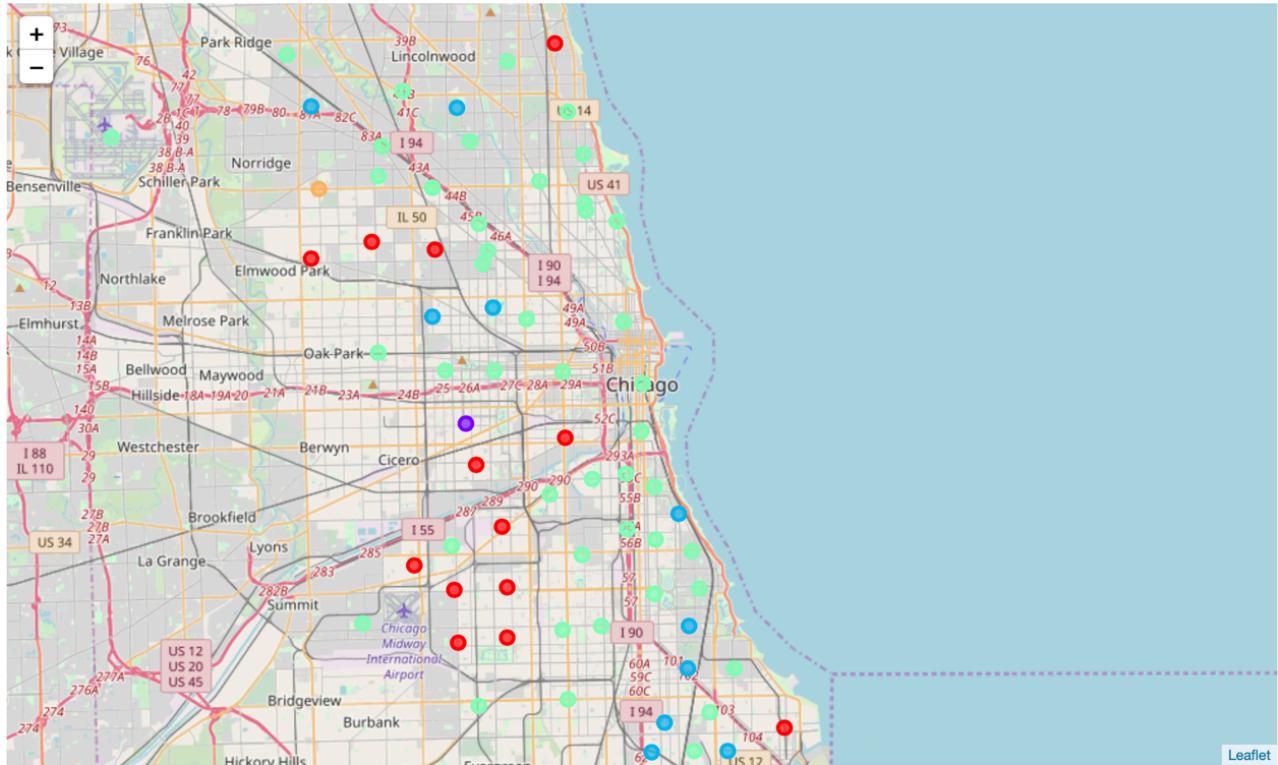
Part 3: Analyzing Data with Machine Learning

In order to understand the underlying relationship between venues and communities, we use unsupervised machine learning method k-means to cluster the communities into 5 clusters and see if there is anything interesting.

All communities are divided into 5 clusters, namely cluster 0, 1, 2, 3, 4. The cluster column is circled in red:

	CommunityArea	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue
0	Albany Park	Albany Park, Mayfair, North Mayfair, Ravenswo...	41.9719	-87.7162	3.0	Sandwich Place	Pizza Place	Donut Shop	Fried Chicken Joint	Bakery	Mobile Phone Shop	Gas Station
1	Archer Heights	Archer Heights	41.8114	-87.7262	3.0	Mexican Restaurant	Mobile Phone Shop	Grocery Store	Gas Station	Optical Shop	Candy Store	Big Box Store
2	Armour Square	Armour Square, Chinatown, Wentworth Gardens	41.84	-87.6331	3.0	Chinese Restaurant	Sports Bar	Hot Dog Joint	Breakfast Spot	Gas Station	Italian Restaurant	Asian Restaurant
3	Ashburn	Ashburn, Ashburn Estates, Beverly View, Crest...	41.7475	-87.7112	3.0	Construction & Landscaping	Cosmetics Shop	Light Rail Station	Italian Restaurant	Automotive Shop	Cuban Restaurant	Currency Exchange
4	Auburn Gresham	Auburn Gresham, Gresham	41.7505	-87.6643	3.0	Fast Food Restaurant	Lounge	Greek Restaurant	Pharmacy	Cosmetics Shop	Electronics Store	Ethiopian Restaurant

Now we create the Chicago map again with each clusters showing with different colors.



Chicago is an interesting city because it is distributed evenly from the very downtown to suburb.

You can clearly see the green clusters are closer to the downtown, while red and blue clusters are in the suburb area.

What's even more interesting is that based on my knowledge of the city of Chicago, the red clusters are relatively poor residential communities, while the other clusters like green (closest to downtown) and blue are residential communities with above average income families. We will discuss this more in the next part.

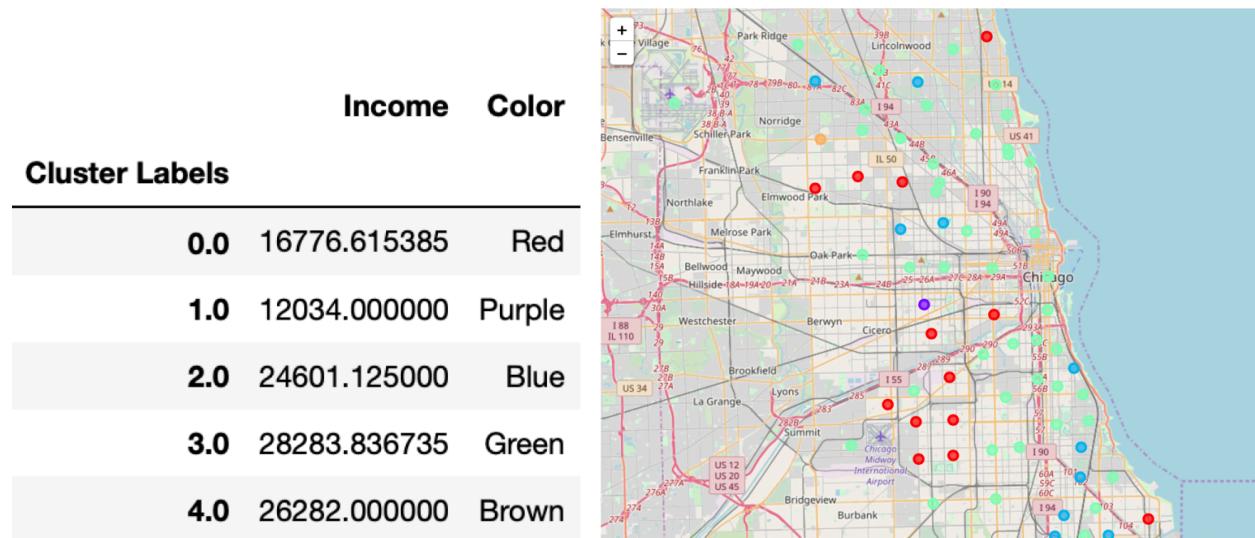
Part 4: Get Our Dataset No.2

It is hard to make that statement without data. So let's look at the data.¶

I got the data from Chicago Data Portal (<https://data.cityofchicago.org/Health-Human-Services/Per-Capita-Income/r6ad-wvtk>). This data is imported and shown below.

Community Area Number	Community Area Name	Percent of Housing Crowded	Percent Households Below Poverty	Percent 16+ Unemployed	Percent Aged 25+ Without High School Diploma	Percent Aged Under 18 or Over 64	Per Capita Income	Hardship Index
0	1.0 Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39.0
1	2.0 West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46.0
2	3.0 Uptown	3.8	24.0	8.9	11.8	22.2	35787	20.0
3	4.0 Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17.0
4	5.0 North Center	0.3	7.5	5.2	4.5	26.2	57123	6.0

After some data processing, we got the average Income of each cluster. Let's also add the color of each cluster.



Part 5: Discussions

So the dataframe shown above agrees with my earlier assumptions. The red clusters are relatively poor residential communities, while the other clusters like green (closest to downtown) and blue are residential communities with above average income families. The Brown cluster only has one community, it is most likely very rich community. The Purple cluster only has one community, it is most likely very poor community.

This is very interesting because we got the clusters from venues near each community. How would the venues near each community correlated to average income? Are certain stores/restaurant open more often in rich communities or vice versa?

In order to understand this question, we find the most three counted venues in each cluster.

Cluster Labels	Income	Color	Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0.0	16776.615385	Red	Mexican Restaurant	Mexican Restaurant	American Restaurant
1.0	12034.000000	Purple	Convenience Store	Yoga Studio	Donut Shop
2.0	24601.125000	Blue	Bus Station	Gymnastics Gym	Park
3.0	28283.836735	Green	Fast Food Restaurant	Pizza Place	Train Station
4.0	26282.000000	Brown	Deli / Bodega	Bar	Yoga Studio

In the above dataframe, we can see that the cluster 0, which is a below average income community cluster, has a lot of Mexican Restaurant. This indicates that these areas, which are most likely Mexican resident areas, are relatively poor.

Cluster 1 only has one community in there, so it may not represent too much. It has Convenience Store in the first place and Yoga Studio at the second place. Notice that Yoga Studio also exist as the third place in cluster 4, which also only has one community.

In cluster 3, which is near downtown area, we see a lot of Fast Food Restaurant and Pizza Place, that's because this is an area mainly lived by single people who live in expensive condos and work in downtown area. They either don't have a kitchen in the apartment or don't have time to cook. So fast food restaurants and pizza places are popular.

Blue Cluster 2 is close to downtown but not that close. This is the area where above average income people with family live in. Many of them still work in the downtown center, but they buy houses in the area where it is cheaper than downtown, so they can buy a bigger house. They still

need public transportsations to go to work, that explains why Bus Station is very common in these areas. A lot of Gymnastics Gym indicates the people live here live healthier and maybe happier. There are many parks in this area, which indicates this is a good area to live in.

Part 6: Ending/Conclusion

So now we can help people who is looking to open a restaurant, and who is trying to start their own business. For example, if someone want to start a Mexican Restaurant, you'd better do it in one of the communities in cluster 0. But try to avoid the communities with too many Mexican Restaurant, because there may be a lot of competition. And if someone wants to open a gym, we could recommend to do it in one of the communities in cluster 2 because the people in this area are more likely to visit a gym from time to time.

Thanks for reading!!!